Spectral Analysis of Symmetric and Anti-Symmetric Pairwise Kernels

Tapio Pahikkala[1], Markus Viljanen[1], Antti Airola[1], and Willem Waegeman[2]

[1]*Department of Information Technology, University of Turku, Joukahaisenkatu 3-5 B, FIN-20520, Turku, Finland, firstname.surname@utu.fi*
[2]*Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, B-9000 Ghent, Belgium, firstname.surname@UGent.be*

June 22, 2015

Abstract

We consider the problem of learning regression functions from pairwise data when there exists prior knowledge that the relation to be learned is symmetric or anti-symmetric. Such prior knowledge is commonly enforced by symmetrizing or anti-symmetrizing pairwise kernel functions. Through spectral analysis, we show that these transformations reduce the kernel's effective dimension. Further, we provide an analysis of the approximation properties of the resulting kernels, and bound the regularization bias of the kernels in terms of the corresponding bias of the original kernel.

1 Introduction

Many real-world phenomena can be described in tems of pairwise relationships between entities. When learning pairwise relations, symmetry and anti-symmetry are two types of prior knowledge constraints that commonly appear when both of the objects in a pair belong to the same domain. A typical example of an application where relationships are often assumed to be symmetric is the prediction of protein-protein interactions: if protein A interacts with protein B, then conversely it also holds that B interacts with A. Typical example of an anti-symmetric relation would be a preference relation: if A is preferred over B, then conversely B is not preferred over A. Commonly used symmetric pairwise kernels include the symmetrized Kronecker [Ben-Hur and Noble, 2005] and Cartesian [Kashima et al., 2009], as well as the metric learning [Vert et al., 2007] kernels. Such kernels are analyzed in more detail by Brunner et al. [2012]. Typical examples of anti-symmetric kernels are the transitive kernel of [Herbrich et al., 2000] used for learning to rank, and the anti-symmetric Kronecker product kernel [Pahikkala et al., 2010] for learning intransitive preference relations.

Kernel-based learning algorithms are some of the most successful learning methods in practise and they also enjoy strong theoretical properties. It is well known in the machine learning literature that the eigenvalues and eigenfunctions of the integral operator of the kernel play a central role in obtaining error estimates in learning theory. One of the most intensively studied quantities depending on the eigenvalues is the so-called *effective dimension* of the kernel, which has since its introduction by Zhang [2002] been used by several other authors [Mendelson, 2003, Caponnetto and De Vito, 2007]. For a recent summary of these results, see Hsu et al. [2014] and references therein. Therefore, the determination of the operator's eigensystem is important in its own right. Another important tool for analysis is the theory of universal kernels pioneered by Steinwart [2002], which indicates that if a kernel has the so-called unversality property, the corresponding hypothesis space can approximate any continuous function arbitrarily well.

Intuitively it seems plausible that enforcing prior knowledge about symmetry or anti-symmetry should result in better generalization, and many promising experimental results have been obtained in the literature (see previous references). However, thus far rigorous theoretical analysisis of the effects that enforcing these properties on the kernel function has on learning has been missing in the literature. As a step towards this direction Waegeman et al. [2012] have shown that when symmetrizing or anti-symmetrizing pairwise kernels that are formed by taking the Kronecker product of two universal kernels, the resulting kernel allows approximating arbitrarily well any symmetric or anti-symmetric continuous function. While these results show that symmetrization or anti-symmetrization does not sacrifice expressive power needed for learning, the results concern only Kronecker product kernels, and do not provide any guarantees that learning would be more efficient with the transformed kernels.

Following are the main contributions and results of our paper:

- The effective dimension of both the symmetrized and anti-symmetrized versions of a pairwise kernel are smaller than that of the original pairwise kernel (see Theorem 4.3).
- The approximation properties of the symmetric and anti-symmetric kernels are analysed (see Theorem 4.6).
- We bound the regularization bias of the symmetric and anti-symmetric kernels in terms of the regularization bias of the original kernel (see Theorem 4.9).

2 Preliminaries

Definition 2.1 (Kernel function)**.** *For any set $\mathcal{X}$, the function K is a kernel if it can be written as the following type of an inner product:*

$$K(x, \overline{x}) = \langle \Phi(x), \Phi(\overline{x}) \rangle \ ,$$

where

$$\Phi : \mathcal{X} \to \mathcal{H}_\Phi$$

is a mapping from $\mathcal{X}$ to a Hilbert space $\mathcal{H}_\Phi$, popularly called the feature space in the literature. Conversely, any kernel can be written as the above type of an inner product. However, neither the feature mapping nor the feature space are unique.

To simplify the forthcoming considerations, we make a couple of extra assumptions of the input space and kernels. Namely, we assume that the input space $\mathcal{X}$ is compact (e.g. closed and bounded) and the kernel functions considered in this article are continuous. Let μ be a probability distribution over $\mathcal{X}$ generating the data. We also assume that μ is a probability density with respect to a Lebesque measure (e.g. we can write $\int_{\mathcal{X}} h(x)d\mu(x) = \int_{\mathcal{X}} h(x)\mu(x)dx$ for any function h).

We make use of the Hilbert space $L^2(\mathcal{X},\mu)$ of square integrable functions on $(\mathcal{X},\mu)$ with the inner product $\langle h,g\rangle_{L^2(\mathcal{X},\mu)} = \int_x h(x)g(x)d\mu(x)$. The elements of the space $L^2(\mathcal{X},\mu)$ are equivalence classes of functions rather that individual functions but this technical detail has no effect on the considerations below.

Definition 2.2 ([Aronszajn, 1950])**.** *For each real-valued kernel K and an input space $\mathcal{X}$, there exists a unique Hilbert space $\mathcal{H}(K)$ known as the reproducing kernel Hilbert space (RKHS):*

1. $K_x \in \mathcal{H}(K) \quad \forall x \in \mathcal{X}$*, where*

$$K_x : \mathcal{X} \to \mathbb{R}$$

   are functions such that $K_x(\overline{x}) = K(x,\overline{x})$

2. span$(\{K_x\}_{x\in\mathcal{X}})$ *is dense in $\mathcal{H}(K)$*

3. *The inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}(K)}$ associated with $\mathcal{H}(K)$ satisfies:*

$$f(x) = \langle f, K_x\rangle \quad \forall f \in \mathcal{H}(K), \quad x \in \mathcal{X}$$

   which is known as the reproducing property. In particular,

$$K(x,\overline{x}) = \langle K_x, K_{\overline{x}}\rangle \quad \forall x,\overline{x} \in \mathcal{X}\ .$$

In the literature, the mapping:

$$\Phi_K : x \to K_x \in \mathcal{H}(K)$$

is often referred to as the canonical feature map of the kernel.

Definition 2.3 (Integral operator of a kernel)**.** *The probability distribution μ over $\mathcal{X}$ yields a linear operator*

$$\mathbf{U}_K : L^2(\mathcal{X},\mu) \to \mathcal{H}(K)$$

defined as

$$\mathbf{U}_K h = \int_{\mathcal{X}} K_x h(x) d\mu(x)\ .$$

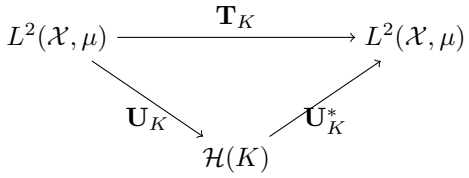
The adjoint of this operator is the inclusion $\mathbf{U}_K^* : \mathcal{H}(K) \hookrightarrow L^2(\mathcal{X},\mu)$, that is,

$$\langle \mathbf{U}_K h, g\rangle_{\mathcal{H}(K)} = \langle h, \mathbf{U}_K^* g\rangle_{L^2(\mathcal{X},\mu)}\ . \tag{1}$$

Note the RKHS norm on the left hand side, determined by the reproducing property, being changed to the $L^2(\mathcal{X},\mu)$ norm on the right. The composition of $\mathbf{U}_K$ with its adjoint is the operator:

$$\mathbf{T}_K : L^2(\mathcal{X},\mu) \to L^2(\mathcal{X},\mu)\ .$$

for all $h \in L^2(\mathcal{X},\mu)$. This decomposition is illustrated in the following commutative diagram:



The operator $\mathbf{T}_K$ can be shown to be continuous, self-adjoint and Hilbert-Schmidt, the last property indicating that its eigenvalues are square-summable, which is characterized below in more detail. We next recollect some classical results from functional analysis required in the forthcoming considerations.

Theorem 2.4 (Spectral theorem for compact operators)**.** *Suppose $\mathcal{L}$ is a Hilbert space and $\mathbf{T} : \mathcal{L} \to \mathcal{L}$ is compact and self-adjoint linear operator. Then, $\mathcal{L}$ has an orthonormal basis $\{\phi_i\}_i$ consisting of eigenvectors of $\mathbf{T}$.*

To compress the forthcoming notation and to take advantage the machinery of operator algebra, we use the following expression for the eigen decomposition of the integral operators:

$$\mathbf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^* \ ,$$

where $\mathbf{V} : e_i \mapsto \phi_i$ and $\mathbf{\Lambda} : e_i \mapsto \lambda_i e_i$, with e_i being the standard basis vectors of l^2.

For the integral operators of continuous kernels on compact domains, we have the following result known as Mercer's theorem:

Theorem 2.5 (Mercer 1909)**.** *Suppose K is a continuous symmetric non-negative definite kernel. Then there is an orthonormal basis $\{\phi_i\}_i$ of $L^2(\mathcal{X})$ consisting of eigenfunctions of T_K such that the corresponding sequence of eigenvalues $\{\lambda_i\}_i$ is nonnegative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on $\mathcal{X}$ and K has the representation*

$$K(x,\overline{x}) = \sum_{j\in\mathbb{N}} \lambda_j \, \phi_j(x) \, \phi_j(\overline{x})$$

where the convergence is absolute and uniform.

The spectral theorem also yields the following corollary about commuting compact and self-adjoint operators sharing the same eigen system (see e.g. Zimmer [1990]):

Corollary 2.6. *Let $\mathcal{T}$ be a Hilbert space and let $\mathbf{T}_1 : \mathcal{L} \to \mathcal{L}$ and $\mathbf{T}_2 : \mathcal{L} \to \mathcal{L}$ be compact and self-adjoint operators, such that $\mathbf{T}_1\mathbf{T}_2 = \mathbf{T}_2\mathbf{T}_1$. Then there is an orthonormal basis $\{\phi_j\}$ of $\mathcal{L}$ such that ϕ_j an eigenvector for both $\mathbf{T}_1$ and $\mathbf{T}_2$.*

Next, we define the concept of majorization for sequences of infinite lengths (see e.g. Li and Busch [2013] and references therein).

Definition 2.7 (Majorization)**.** *Let $\mathbf{r} = (r_i)_{i=1}^{\infty} \in c_0^*$ and $\mathbf{s} = (s_i)_{i=1}^{\infty} \in c_0^*$ where c_0^* is the positive cone of sequences decreasing monotonically to 0. We say that $\mathbf{s}$ majorizes $\mathbf{r}$, denoted as $\mathbf{r} \prec \mathbf{s}$ if*

$$\sum_{i=1}^{m} r_i \leq \sum_{i=1}^{m} s_i \ \forall m \in \mathbb{N} \text{ and } \sum_{i=1}^{\infty} r_i = \sum_{i=1}^{\infty} s_i \ .$$

In particular, for two trace class operators $\mathbf{T}_1$ and $\mathbf{T}_2$ on a Hilbert space, we say that $\mathbf{T}_2 \prec \mathbf{T}_1$ if the sequence of eigenvalues of $\mathbf{T}_1$ majorizes the sequence of eigenvalues of $\mathbf{T}_2$.

The next result is a recent generalization by Li and Busch [2013] of the classical Uhlmann's theorem for infinite dimensional Hilbert spaces. Before that, we also define the doubly-stochastic operations, which is also by Li and Busch [2013]:

Definition 2.8 (Doubly-stochastic operation)**.** *Let $\mathcal{T}(\mathcal{L})$ denote the (Banach) space of all trace class operators on a Hilbert space $\mathcal{L}$. We say that operation $\Gamma : \mathcal{T}(\mathcal{L}) \to \mathcal{T}(\mathcal{L})$ is doubly-stochastic if it preserves trace (e.g. $\mathrm{trace}(\mathbf{T}) = \mathrm{trace}(\Gamma(\mathbf{T}))$), is unital indicating that $\mathbf{I} = \Gamma(\mathbf{I})$ for the identity operator $\mathbf{I}$ on the Hilbert space, and there exists a sequence $\{\mathbf{E}_i\}_{i=1}^{\infty}$ of compact operators on the Hilbert space $\mathcal{L}$, known in the literature as the Kraus operators, such that the operation can be written as*

$$\Gamma(\mathbf{T}) = \sum_{i=1}^{\infty} \mathbf{E}_i \mathbf{T} \mathbf{E}_i^* \ . \tag{2}$$

Theorem 2.9 (Uhlmann's theorem for infinite dimensional Hilbert spaces)**.** *If $\mathbf{T}_1$ and $\mathbf{T}_2$ are trace-class operators on a Hilbert space, then $\mathbf{T}_2 \prec \mathbf{T}_1$ iff there exists a doubly-stochastic operation Γ such that $\mathbf{T}_2 = \Gamma(\mathbf{T}_1)$.*

3 Pairwise Kernels

Let us next define the family of pairwise kernels. Assume that the input space can be written as

$$\mathcal{X} = \mathcal{P}^2$$

where $\mathcal{P}$ is a compact metric space. The kernels over $\mathcal{P}^2$ can accordingly be written as the following types of inner products

$$K(v, v', \overline{v}, \overline{v}') = \langle \Phi(v, v'), \Phi(\overline{v}, \overline{v}') \rangle \ ,$$

where $v, v', \overline{v}, \overline{v}' \in \mathcal{P}$ and Φ is a joint feature mapping over a pair of inputs, that is, $\Phi(v, v')$ is a feature space representation for an ordered pair (v, v').

Next, we define certain specific types of pairwise kernels, starting from the permuted kernel:

Definition 3.1 (Permuted pairwise kernel)**.** *Let $K(v, v', \overline{v}, \overline{v}')$ be an arbitrary kernel on $\mathcal{P}^2$. Then, its permuted pairwise kernel is*

$$K^P(v, v', \overline{v}, \overline{v}') = K(v', v, \overline{v}', \overline{v}) \ .$$

An immediate step forward is to define the following type of kernels that are invariant to the permutations in the above defined sense:

Definition 3.2 (Permutation invariant pairwise kernels)**.** *We say that a kernel $K^{PI}(v, v', \overline{v}, \overline{v}')$ on $\mathcal{P}^2$ is permutation invariant if it is equal to its permuted kernel, that is,*

$$K^{PI}(v, v', \overline{v}, \overline{v}') = K^{PI}(v', v, \overline{v}', \overline{v}) \ .$$

A natural way to construct a permutation invariant kernel from a given pairwise kernel K is to consider the projection from the set of all kernels to the set of permutation invariant kernels:

$$K^{PI}(v, v', \overline{v}, \overline{v}') = \frac{1}{2} \left(K(v, v', \overline{v}, \overline{v}') + K(v', v, \overline{v}', \overline{v}) \right) \ .$$

Our next step is to define the well-known symmetric pairwise kernels as well as their anti-symmetric counterparts:

Definition 3.3 (Symmetric and anti-symmetric pairwise kernels)**.** *We say that a kernel $K^S(v, v', \overline{v}, \overline{v}')$ on $\mathcal{P}^2$ is a symmetric pairwise kernel if*

$$K^S(v, v', \overline{v}, \overline{v}') = K^S(v', v, \overline{v}, \overline{v}') \ .$$

Analogously, we say that a kernel $K^A(v, v', \overline{v}, \overline{v}')$ on $\mathcal{P}^2$ is an anti-symmetric pairwise kernel if

$$K^A(v, v', \overline{v}, \overline{v}') = -K^A(v', v, \overline{v}, \overline{v}') \ .$$

Similarly to the permutation invariance, one can construct symmetric and anti-symmetric kernels from an arbitrary kernel $K(v, v', \overline{v}, \overline{v}')$ with the following projections:
$K^S(v, v', \overline{v}, \overline{v}') =$

$$\frac{1}{4}\Big(K(v, v', \overline{v}, \overline{v}') + K(v', v, \overline{v}, \overline{v}') + K(v, v', \overline{v}', \overline{v}) + K(v', v, \overline{v}', \overline{v})\Big)$$

and $K^A(v, v', \overline{v}, \overline{v}') =$

$$\frac{1}{4}\Big(K(v, v', \overline{v}, \overline{v}') - K(v', v, \overline{v}, \overline{v}') - K(v, v', \overline{v}', \overline{v}) + K(v', v, \overline{v}', \overline{v})\Big) \ ,$$

respectively.

The following connection between the symmetric, anti-symmetric and permutation invariant kernels is immediate:

Lemma 3.4. *Both the symmetric and anti-symmetric pairwise kernels are permutation invariant. Moreover, if $K^S(v, v', \overline{v}, \overline{v}')$ and $K^A(v, v', \overline{v}, \overline{v}')$ are the symmetric and anti-symmetric forms of a kernel $K(v, v', \overline{v}, \overline{v}')$ obtained with the projections given in Definition 3.3, then the permutation invariant form of the kernel obtained with the projection given in Definition 3.2 can be expressed as the sum of the symmetric and and anti-symmetric forms:*

$$K^{PI}(v, v', \overline{v}, \overline{v}') = K^S(v, v', \overline{v}, \overline{v}') + K^A(v, v', \overline{v}, \overline{v}') \ .$$

■

3.1 Spectral Analysis of Pairwise Kernels

We next study the relationship between the integral operators of the permutation invariant, symmetric and anti-symmetric kernels to the corresponding integral operator of the original kernel they were constructed from.

Theorem 3.5. *Let $K(v, v', \overline{v}, \overline{v}')$ be an arbitrary pairwise kernel and let K^{PI}, K^S and K^A be its permutation invariant, symmetric and anti-symmetric forms. Moreover, let $\mathbf{T}_K$, $\mathbf{T}_{K^{PI}}$, $\mathbf{T}_{K^S}$ and $\mathbf{T}_{K^A}$ be the integral operators of the kernels K, K^{PI}, K^S and K^A, respectively. Then,*

$$\begin{aligned}
\mathbf{T}_{K^P} &= \mathbf{P}^{\mu *}\mathbf{T}_K\mathbf{P}^{\mu} \\
\mathbf{T}_{K^S} &= \mathbf{S}^{\mu *}\mathbf{T}_K\mathbf{S}^{\mu} \\
\mathbf{T}_{K^A} &= \mathbf{A}^{\mu *}\mathbf{T}_K\mathbf{A}^{\mu} \\
\mathbf{T}_{K^{PI}} &= \frac{1}{2}\left(\mathbf{T}_K + \mathbf{P}^{\mu *}\mathbf{T}_K\mathbf{P}^{\mu}\right) \qquad (3) \\
&= \mathbf{S}^{\mu *}\mathbf{T}_K\mathbf{S}^{\mu} + \mathbf{A}^{\mu *}\mathbf{T}_K\mathbf{A}^{\mu} \ , \qquad (4)
\end{aligned}$$

where

$$\begin{aligned}\mathbf{P}^{\mu} :& L^2(\mathcal{P}^2, \mu) \to L^2(\mathcal{P}^2, \mu) \\ & h(\overline{v}, \overline{v}') \mapsto \frac{\mu(\overline{v}', \overline{v})}{\mu(\overline{v}, \overline{v}')} h(\overline{v}', \overline{v})\end{aligned}$$

is an operator to which we refer as the permutation operator with respect to the measure μ, and whose adjoint is

$$h(\overline{v}, \overline{v}') \mapsto \frac{\mu(\overline{v}, \overline{v}')}{\mu(\overline{v}', \overline{v})} h(\overline{v}', \overline{v}) \ ,$$

and

$$\begin{aligned}\mathbf{S}^{\mu} =& \frac{1}{2} \left(\mathbf{I} + \mathbf{P}^{\mu} \right) \\ \mathbf{A}^{\mu} =& \frac{1}{2} \left(\mathbf{I} - \mathbf{P}^{\mu} \right)\end{aligned}$$

are projection operators to which we refer as the symmetrizer and anti-symmetrizer with respect to the measure μ, and $\mathbf{I}$ is the identity operator of $L^2(\mathcal{P}^2, \mu)$.

See Section 5.1 for a proof.

Next, we look on what can be said about the spectrum of the integral operators considered in the above theorem. This consideration can be divided into the important special case of the measure μ being symmetric, that is

$$\mu(\overline{v}, \overline{v}') = \mu(\overline{v}', \overline{v}), \forall (\overline{v}, \overline{v}') \in \mathcal{P}^2$$

and to the general case. The measure is symmetric, for example, in various types of ranking and preference learning tasks as is considered more in detail below. In addition, many other pairwise learning problems with non-symmetric measure can be turned to problems with a symmetric measure by the technique known as virtual examples. That is, whenever a datum (v, v') is drawn from μ, one also introduces a virtual example (v', v) with the same output if the problem is considered to be symmetric or with the opposite output in the anti-symmetric case. With symmetric μ, the symmetrizer and anti-symmetrizer projections do not depend on the measure and we denote them simply as $\mathbf{S}$ and $\mathbf{A}$.

Corollary 3.6. *If λ_i^K, $\lambda_i^{K^S}$ and $\lambda_i^{K^A}$ denote the eigenvalues of $\mathbf{T}_K$, $\mathbf{T}_{K^S}$ and $\mathbf{T}_{K^A}$, respectively, then*

$$\lambda_i^{K^S} \leq \lambda_i^K \text{ and } \lambda_i^{K^A} \leq \lambda_i^K \text{ for } i = 1, 2, \ldots \tag{5}$$

If μ is symmetric, the set of operators $\{\mathbf{S}, \mathbf{A}, \mathbf{T}_{K^S}, \mathbf{T}_{K^A}, \mathbf{T}_{K^{PI}}\}$ commutes, which in turn indicates that they can be diagonalized simultaneously as follows:

$$\begin{aligned}\mathbf{T}^{PI} &= \mathbf{V} \mathbf{\Lambda}^{PI} \mathbf{V}^* \ , \\ \mathbf{T}^{S} &= \mathbf{V} \mathbf{\Lambda}^{S} \mathbf{V}^* \ , \\ \mathbf{T}^{A} &= \mathbf{V} \mathbf{\Lambda}^{A} \mathbf{V}^* \ , \\ \mathbf{S} &= \mathbf{V} \mathbf{I}^{S} \mathbf{V}^* \ , \\ \mathbf{A} &= \mathbf{V} \mathbf{I}^{A} \mathbf{V}^* \ ,\end{aligned}$$

where $\mathbf{V}$ *is an unitary operator containing the eigenfunctions and* $\mathbf{\Lambda}^{PI}$, $\mathbf{\Lambda}^{S}$, $\mathbf{\Lambda}^{A}$, $\mathbf{I}^{S}$ *and* $\mathbf{I}^{A}$ *are operators containing the corresponding eigenvalues of the five operators under consideration, and*

$$\mathbf{\Lambda}^{PI} = \mathbf{\Lambda}^{S} + \mathbf{\Lambda}^{A} \ , \tag{6}$$

if the eigenvalues are arranged in the order determined by the order of eigenfunction in $\mathbf{V}$.

Finally, if μ *is symmetric, then*

$$\mathbf{T}_{K^{PI}} \prec \mathbf{T}_{K} \tag{7}$$

(e.g. the sequence of eigenvalues of $\mathbf{T}_K$ *majorizes the sequence of eigenvalues of* $\mathbf{T}_{K^{PI}}$*).*

Proof. Since $\mathbf{A}^{\mu}$ is a projection matrix, $\mathbf{A}^{\mu}(L^2(\mathcal{P}^2,\mu)) \subset L^2(\mathcal{P}^2,\mu)$, this constrains the action of the integral operator $\mathbf{T}_K$ onto the range of $\mathbf{A}^{\mu}$, which is a subspace of $L^2(\mathcal{P}^2,\mu)$. The eigenfunctions ϕ_i associated with nonzero eigenvalues λ_i of $\mathbf{T}_{K^A}$ belong to this subspace, and satisfy (Aronszajn [1948]):

$$\mathbf{T}_K\phi_i - \lambda_i\phi_i = p \text{ with } p \perp \mathbf{A}^{\mu}(L^2(\mathcal{P}^2,\mu)) \ .$$

Since $\mathbf{A}^{\mu}(L^2(\mathcal{P}^2,\mu)) \subset L^2(\mathcal{P}^2,\mu)$, we can use a well known theorem (see e.g. Aronszajn [1948] and references therein) to obtain:

$$\text{and } \lambda_i^{K^A} \leq \lambda_i^{K} \text{ for } i = 1, 2, \ldots$$

and the case with $\mathbf{T}_{K^S}$ goes analogously.

We observe that, with symmetric μ, the operators $\mathbf{S}$ and $\mathbf{A}$ are self-adjoint, and hence orthogonal projections. Furthermore, they are orthogonal with each other, that is

$$\mathbf{SA} = \mathbf{AS} = 0 \ , \tag{8}$$

and hence the set $\{\mathbf{S}, \mathbf{A}, \mathbf{T}_{K^S}, \mathbf{T}_{K^A}, \mathbf{T}_{K^{PI}}\}$ of operators commutes, and therefore, according to Corollary 2.6, they share the same eigenfunctions.

Finally, (7) follows from the Uhlmann's theorem, since we can define an operation:

$$\begin{aligned} \Gamma :& \mathcal{T}(L^2(\mathcal{P}^2,\mu)) \to \mathcal{T}(L^2(\mathcal{P}^2,\mu)) \\ & \mathbf{T}_K \mapsto \frac{1}{2}\left(\mathbf{T}_K + \mathbf{P}\mathbf{T}_K\mathbf{P}\right) \end{aligned}$$

for which $\mathbf{T}_{K^{PI}} = \Gamma(\mathbf{T}_K)$ and which is doubly stochastic, because it is both trace preserving (as shown above), unital due to $\mathbf{PP} = \mathbf{I}$, and the set of Kraus operators fulfilling (2) is $\{\frac{1}{2}\mathbf{I}, \frac{1}{2}\mathbf{P}\}$. □

It is interesting to note the following observation about the common eigensystem of the operators $\{\mathbf{S}, \mathbf{A}, \mathbf{T}_{K^S}, \mathbf{T}_{K^A}, \mathbf{T}_{K^{PI}}\}$:

Remark 3.7. *All the eigenfunctions of* $\mathbf{T}_{K^{PI}}$ *are either symmetric or anti-symmetric, and the corresponding eigenvalues are cleared to zeros when one applies* $\mathbf{S}$ *or* $\mathbf{A}$*. Since* $\mathbf{S}$ *and* $\mathbf{A}$ *are orthogonal projections, their eigenvalues are either zeros or ones, and the ones in* $\mathbf{S}$ *correspond to the symmetric functions and zeros to the anti-symmetric ones, and vice versa for* $\mathbf{A}$*.*

4 Error Bounds

Let

$$I(f) = \int_{\mathcal{X}\times\mathcal{Y}} L(f(x), y) d\rho(x, y), \tag{9}$$

where L is a loss function, denote the expected risk of f. For the squared loss, the minimizer of (9) is the so-called regression function

$$f^*(x) = \int_{\mathcal{Y}} y d\rho(x, y).$$

The hypothesis spaces under our consideration in this paper do not necessarily include the regression function, and hence another quantity of interest is the error associated to the given RKHS $\mathcal{H}$:

$$\inf_{f\in\mathcal{H}} I(f).$$

If we have a prior knowledge, for example, that the underlying regression function is anti-symmetric, then we can immediately assume that the errors associated to a kernel K and its anti-symmetric counterpart K^A are equal. That is, we do not lose any expressiveness by restricting our hypothesis space to anti-symmetric functions. The next question is whether we can gain anything with the restriction.

Our next quantity of interest is the minimizer $f_{T,\lambda}$ of the regularized empirical risk on a training set T and a regularization parameter λ. In particular, we aim to analyze the effect of using either the permutation-invariant, symmetric, or anti-symmetric forms instead of the original kernel on the discrepancy

$$I(f_{\lambda,T}) - \inf_{f\in\mathcal{H}(K)} I(f)$$

known in the literature as the excess error.

Following Hsu et al. [2014], we split the consideration of the excess error into three parts:

$$I(f_{\lambda,T}) - \inf_{f\in\mathcal{H}(K)} I(f) \leq \epsilon_{rg} + \epsilon_{bs} + \epsilon_{vr} + 2(\sqrt{\epsilon_{rg}\epsilon_{bs}} + \sqrt{\epsilon_{rg}\epsilon_{vr}} + \sqrt{\epsilon_{bs}\epsilon_{vr}}) ,$$

where ϵ_{rg}, ϵ_{bs}, and ϵ_{vr} are, respectively, the bias caused by regularization, the bias caused by the random drawing of the training inputs, and the variance caused by noise in the outputs. We briefly consider each of these in turn in the following subsections.

4.1 Effective Dimension

As discussed by Hsu et al. [2014] and also earlier by many other authors (see e.g. (Zhang [2005], Caponnetto and De Vito [2007]), the variance term ϵ_{vr} can be roughly characterized with a concept known as the effective dimension:

Definition 4.1 (Effective dimension)**.** *The effective dimension $D(K, \mu, \lambda)$ of the kernel K with respect to the measure μ and the regularization parameter value $\lambda > 0$ is defined as:*

$$D(K, \mu, \lambda) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} ,$$

where λ_i are the eigenvalues of the integral operator of the kernel K.

The next result shows that the eigenvalue majorization of the integral operators of kernels is connected to the effective dimension of the kernels:

Proposition 4.2. *Let K_1 and K_2 be kernels, and $\mathbf{T}_1$ and $\mathbf{T}_2$ their integral operators with measure μ, with $\operatorname{trace}(\mathbf{T}_1) = \operatorname{trace}(\mathbf{T}_2)$. Then,*

$$\mathbf{T}_2 \prec \mathbf{T}_1 \Rightarrow D(K_2, \mu, \lambda) > D(K_1, \mu, \lambda) \quad \forall \lambda > 0 \; .$$

Proof. We recollect the following result recently proven by Mari et al. [2014] that extends a well-known result for sequences of infinite lengths. Let $\mathbf{r} = (r_i)_{i=1}^{\infty} \in c_0^*$ and $\mathbf{s} = (s_i)_{i=1}^{\infty} \in c_0^*$ with $\sum_{i=1}^{\infty} r_i = \sum_{i=1}^{\infty} s_i = 1$. Then,

$$\mathbf{r} \prec \mathbf{s} \Leftrightarrow \sum_{i=1}^{\infty} \rho(r_i) \geq \sum_{i=1}^{\infty} \rho(s_i) \; .$$

for all real non-negative strictly concave function ρ defined on the segment $[0, 1]$. The result follows immediately (with scaling the eigenvalues), since $\rho(r) = r/(r + \lambda)$ is real-valued, non-negative and strictly concave for $r, \lambda > 0$. □

Given the above analysis of the eigensystems of the considered pairwise kernels, we end up to the following results about their effective dimensions:

Theorem 4.3. *If K is a pairwise kernel, then*

$$D(K^S, \mu, \lambda) \leq D(K, \mu, \lambda) \tag{10}$$

and

$$D(K^A, \mu, \lambda) \leq D(K, \mu, \lambda) \; . \tag{11}$$

If the measure μ is symmetric, we also have

$$D(K, \mu, \lambda) \leq D(K^{PI}, \mu, \lambda) \; . \tag{12}$$

Proof. The inequalities (10) and (11) follow straightforwardly from (5), and the inequality (12) follows from Corollary 4.2 and (7). □

4.2 Approximation Analysis

We next rurn our attention to the bias caused by the random drawing of the training inputs. According to Hsu et al. [2014], this bias is affected, in addition to the above considered effective dimension and the regularization bias considered below, by the approximation error caused by the hypothesis space being too limited. In contrast, the approximation error is zero if the hypothesis space contains the regression function or functions that can approximate it arbitrarily closely. To guarantee that the hypothesis space is expressive enough to approximate any function, we may use kernels that are universal. On the other hand, if we have prior knowledge about the properties of the regression function, for example, if we know it to be symmetric or anti-symmetric, we may restrict the hypothesis space accordingly.

Related to the bias by random design, we also point out a recent result by Brunner et al. [2012] which shows an equivalence between the use of a symmetric pairwise kernel and the original kernel with a symmetrized training set. We omit its detailed consideration here due to lack of space.

To formalize these concepts, we first recollect the definition of universal kernels.

Definition 4.4 (Steinwart [2002])**.** *A continuous kernel K on a compact metric space $\mathcal{X}$ (i.e. $\mathcal{X}$ is closed and bounded) is called universal if the RKHS induced by K is dense in $C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of all continuous functions $f : \mathcal{X} \to \mathbb{R}$.*

Accordingly, the hypothesis space induced by the kernel K can approximate any function in $C(\mathcal{X})$ arbitrarily well, and hence it is said to have the universal approximating property.

While the universal approximating property guarantees that the RKHS can, in theory, learn any concept, we do not necessarily have a need for it if we have prior knowledge about certain properties of the concept to be learned. Thus, we also define an analogous concept for non-universal kernels:

Definition 4.5. *Let K be a continuous kernel K on a compact metric space $\mathcal{X}$ and let $\mathcal{F} \subseteq C(\mathcal{X})$. If $\mathcal{F} \subseteq \mathcal{H}(K)$, the definition of RKHS indicates that, for every function $f \in C(\mathcal{X})$ and every $\epsilon > 0$, there exists a set of input points $\{x_i\}_{i=1}^m \in \mathcal{X}$ and real numbers $\{\alpha_i\}_{i=1}^m$, with $m \in \mathbb{N}$, such that*

$$\max_{x \in \mathcal{X}} \left\{ \left| f(x) - \sum_{i=1}^{m} \alpha_i K(x_i, x) \right| \right\} \leq \epsilon.$$

Accordingly, the hypothesis space induced by the kernel K can approximate any function in $\mathcal{F}$ arbitrarily well, and hence we say that the RKHS $\mathcal{H}(K)$ can approximate $\mathcal{F}$.

Armed with the above definitions, we present the next result characterizing the approximation properties of the symmetric and anti-symmetric kernels:

Theorem 4.6. *Let $\mathcal{F} \subseteq C(\mathcal{P}^2)$ be an arbitrary set of continuous functions, and let*

$$\mathcal{S} = \{t \mid r \in \mathcal{F}, t(v, v') = r(v, v') + r(v', v)\}$$
$$\mathcal{A} = \{t \mid r \in \mathcal{F}, t(v, v') = r(v, v') - r(v', v)\}$$

be the sets of symmetric and anti-symmetric functions determined by $\mathcal{F}$. Moreover, let $K(v, v', \overline{v}, \overline{v}')$ be a kernel on $\mathcal{P}^2$ and let $K^S(v, v', \overline{v}, \overline{v}')$ and $K^A(v, v', \overline{v}, \overline{v}')$ be the corresponding symmetric and anti-symmetric kernels. If $\mathcal{F} \subseteq \mathcal{H}(K)$, then $\mathcal{S} \subseteq \mathcal{H}\left(K^S\right)$ and $\mathcal{A} \subseteq \mathcal{H}\left(K^A\right)$.

See Section 5.2 for a proof. This theorem is a generalization of the result of Waegeman et al. [2012], who proved that this result holds for the special cases of the symmetric and anti-symmetric Kronecker product kernel.

As an example of an anti-symmetric kernel popularly used in the machine learning literature, we may consider the following one originally analyzed by [Herbrich et al., 2000]. Given a base kernel $K^{\mathcal{P}}(v, \overline{v})$ over the objects, the pairwise learning to rank approach corresponds to using the following transitive pairwise kernel:

$$\frac{1}{4}\left(K^{\mathcal{P}}(v, \overline{v}) - K^{\mathcal{P}}(v', \overline{v}) - K^{\mathcal{P}}(v, \overline{v}') + K^{\mathcal{P}}(v', \overline{v}')\right)$$

In the theoretical framework considered in this paper, this kernel can be interpreted as the anti-symmetrization of the pointwise kernel $K(v, v', \overline{v}, \overline{v}') = K^{\mathcal{P}}(v, \overline{v})$, that simply ignores the second pair. The approximation properties of this kernel are thus formalized in the following corollary:

Corollary 4.7. *Let*

$$\mathcal{R} = \left\{ t \mid t \in C(\mathcal{P}^2), \exists r \in C(\mathcal{P}), t(v, v') = r(v) - r(v') \right\}$$

be the set of all continuous ranking functions from $\mathcal{P}^2$ to $\mathbb{R}$. If $K^{\mathcal{P}}(v, \overline{v})$ on $\mathcal{P}$ is universal, then the RKHS of the transitive kernel [Herbrich et al., 2000] defined as $K_T(v, v', \overline{v}, \overline{v}') =$

$$\frac{1}{4}\left(K^{\mathcal{P}}(v, \overline{v}) - K^{\mathcal{P}}(v', \overline{v}) - K^{\mathcal{P}}(v, \overline{v}') + K^{\mathcal{P}}(v', \overline{v}')\right) \tag{13}$$

can approximate $\mathcal{R}$.

Proof. We select

$$\mathcal{F} = \left\{ f \mid f \in C(\mathcal{P}^2), \exists r \in C(\mathcal{P}), t(v, v') = r(v) \right\}$$

and apply Theorem 4.6. □

4.3 Regularization Bias

The following expression of the bias caused by regularization is known in the literature (see e.g. Hsu et al. [2014]) but we show it here for the completeness, because we express it in somewhat different form.

Lemma 4.8. *Let f be the regression function and $\mathbf{T}_K$ the integral operator of a kernel K. Further, let h^λ be the minimizer of the regularized mean squared error*

$$\int_{\mathcal{X}} \left(f - \mathbf{U}_K^* h\right)^2 d\mu + \lambda \|h\|_{\mathcal{H}(K)} \,, \tag{14}$$

and let $f^\lambda = \mathbf{U}_K^* h^\lambda$. Then, f^λ can be expressed as

$$f^\lambda = \mathbf{V}\mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I})^{-1}\mathbf{V}^* f \,,$$

and the bias caused by regularization as

$$\epsilon_{rg}(f, \mathbf{T}_K, \lambda) = \lambda^2 \left\langle f, \left(\mathbf{T}_K + \lambda \mathbf{I}\right)^{-2} f \right\rangle \,,$$

where $\mathbf{T}_K = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ is the eigen decomposition of $\mathbf{T}_K$, and the operator-vector products are in $L^2(\mathcal{X}, \mu)$.

See Section 5.3 for a proof.

Interestingly, if the same value of the regularization parameter is used for both the original kernel and its permutation invariant, symmetric or anti-symmetric forms, the type depending on the prior knowledge we have about the regression function, the regularization bias may get worse even if we use the correct type of modification of the kernel. In fact, one can find examples of symmetric regression functions for which the kernel symmetrization decreases the bias and other symmetric regression functions for which the bias is increased. However, the increase or decrease of the bias is rather mild and it is characterized by the following result:

Theorem 4.9. *Let us assume* $K \max_{(v,v')\in\mathcal{P}^2} K(v,v',v,v') = 1$. *This assumption can be done without losing generality due to the kernels being bounded.*

If the measure μ *is symmetric and the regression function is symmetric (anti-symmetric), the bias caused by regularization is the same for the kernels* K^{PI} *and* K^S (K^{PI} *and* K^A*) with all values of* λ*. Moreover, the bias caused by regularization with the amount* λ *for the kernel* K *and* K^{PI} *has the following relationship:*

$$\left(1 - \frac{(\lambda^2 - (\lambda+1)^2)^2}{(\lambda^2 + (\lambda+1)^2)^2}\right) \epsilon_{rg}(f, \mathbf{T}_K, \lambda) \leq \epsilon_{rg}(f, \mathbf{T}_{K^{PI}}, \lambda) \leq \left(1 + \frac{1}{4\lambda^2 + 4\lambda}\right) \epsilon_{rg}(f, \mathbf{T}_K, \lambda) \, .$$

See Section 5.4 for a proof.

5 Proofs

5.1 Proof of Theorem 3.5

Proof. We begin by considering the integral operator of the anti-symmetric kernel. For $h, g \in L^2(\mathcal{P}^2, \mu)$,

$$\begin{aligned}
\langle \mathbf{T}_{K^A} h, g\rangle_{L^2(\mathcal{P}^2,\mu)} &= \int_{\mathcal{P}^2} g(v,v) \left(\int_{\mathcal{P}^2} K^A(v,v',\overline{v},\overline{v}') h(\overline{v},\overline{v}') d\mu \right) d\mu \\
&= \int_{\mathcal{P}^2} \int_{\mathcal{P}^2} g(v,v) K^A(v,v',\overline{v},\overline{v}') h(\overline{v},\overline{v}') d\mu d\mu \\
&= \frac{1}{4} \int_{\mathcal{P}^2} \int_{\mathcal{P}^2} g(v,v') K(v,v',\overline{v},\overline{v}') h(\overline{v},\overline{v}') d\mu d\mu \\
&\quad - \frac{1}{4} \int_{\mathcal{P}^2} \int_{\mathcal{P}^2} g(v,v') K(v',v,\overline{v},\overline{v}') h(\overline{v},\overline{v}') d\mu d\mu \\
&\quad - \frac{1}{4} \int_{\mathcal{P}^2} \int_{\mathcal{P}^2} g(v,v') K(v,v',\overline{v}',\overline{v}) h(\overline{v},\overline{v}') d\mu d\mu \\
&\quad + \frac{1}{4} \int_{\mathcal{P}^2} \int_{\mathcal{P}^2} g(v,v') K(v',v,\overline{v}',\overline{v}) h(\overline{v},\overline{v}') d\mu d\mu \\
&= \frac{1}{4} \left\langle \int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') d\mu, \int_{\mathcal{P}^2} K_{(v,v')} g(v,v') d\mu \right\rangle \\
&\quad - \frac{1}{4} \left\langle \int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') d\mu, \int_{\mathcal{P}^2} K_{(v',v)} g(v',v) d\mu \right\rangle \\
&\quad - \frac{1}{4} \left\langle \int_{\mathcal{P}^2} K_{(\overline{v}',\overline{v})} h(\overline{v}',\overline{v}) d\mu, \int_{\mathcal{P}^2} K_{(v,v')} g(v,v') d\mu \right\rangle \\
&\quad + \frac{1}{4} \left\langle \int_{\mathcal{P}^2} K_{(\overline{v}',\overline{v})} h(\overline{v}',\overline{v}) d\mu, \int_{\mathcal{P}^2} K_{(v',v)} g(v',v) d\mu \right\rangle \\
&= \left\langle \int_{\mathcal{P}^2} \Phi^A_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') d\mu, \int_{\mathcal{P}^2} \Phi^A_{(v,v')} g(v,v') d\mu \right\rangle \, ,
\end{aligned}$$

where $\Phi^A_{(\overline{v},\overline{v}')} = \frac{1}{2}\left(K_{(\overline{v},\overline{v}')} - K_{(\overline{v}',\overline{v})}\right)$. Then,

$$
\begin{aligned}
&\int_{\mathcal{P}^2} \Phi^A_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') d\mu(\overline{v},\overline{v}') \\
&= \frac{1}{2}\int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') d\mu(\overline{v},\overline{v}') - \frac{1}{2}\int_{\mathcal{P}^2} K_{(\overline{v}',\overline{v})} h(\overline{v},\overline{v}') d\mu(\overline{v},\overline{v}') \\
&= \frac{1}{2}\int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} h(\overline{v},\overline{v}') \mu(\overline{v},\overline{v}') d(\overline{v},\overline{v}') - \frac{1}{2}\int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} h(\overline{v}',\overline{v}) \mu(\overline{v}',\overline{v}) d(\overline{v},\overline{v}') \\
&= \frac{1}{2}\int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} \left(h(\overline{v},\overline{v}') - \frac{\mu(\overline{v}',\overline{v})}{\mu(\overline{v},\overline{v}')} h(\overline{v}',\overline{v}) \right) \mu(\overline{v},\overline{v}') d(\overline{v},\overline{v}') \\
&= \int_{\mathcal{P}^2} K_{(\overline{v},\overline{v}')} \left(\mathbf{A}^\mu h(\overline{v},\overline{v}') \right) d\mu(\overline{v},\overline{v}') \\
&= \mathbf{U}_K(\mathbf{A}^\mu h)
\end{aligned}
$$

Accordingly, we observe that:

$$
\begin{aligned}
\langle \mathbf{T}_{K^A} h, g\rangle_{L^2(\mathcal{P}^2,\mu)} =& \langle \mathbf{U}_K \mathbf{A}^\mu h, \mathbf{U}_K \mathbf{A}^\mu g\rangle_{\mathcal{H}} \\
=& \langle \mathbf{A}^\mu h, \mathbf{U}_K^* \mathbf{U}_K \mathbf{A}^\mu g\rangle_{L^2(\mathcal{P}^2,\mu)} \\
=& \langle h, \mathbf{A}^{\mu *} \mathbf{U}_K^* \mathbf{U}_K \mathbf{A}^\mu g\rangle_{L^2(\mathcal{P}^2,\mu)} \\
=& \langle h, \mathbf{A}^{\mu *} \mathbf{T}_K \mathbf{A}^\mu g\rangle_{L^2(\mathcal{P}^2,\mu)} ,
\end{aligned}
$$

that is, the integral operator of the anti-symmetric kernel is $\mathbf{T}_{K^A} = \mathbf{A}^{\mu *} \mathbf{T}_K \mathbf{A}^\mu$.

The integral operators of the other kernels can be constructed analogously via the feature mappings:

$$
\begin{aligned}
\Phi_{K^P} h =& \mathbf{U}_K(\mathbf{P}^\mu h) \\
\Phi_{K^S} h =& \mathbf{U}_K(\mathbf{S}^\mu h) \\
\Phi_{K^{PI}} h =& \mathbf{U}_K \left(\begin{pmatrix} \mathbf{I} \\ \mathbf{P}^\mu \end{pmatrix} h \right) ,
\end{aligned}
$$

where $\begin{pmatrix} \mathbf{I} \\ \mathbf{P}^\mu \end{pmatrix}$ is the operator obtained by stacking the operators $\mathbf{I}$ and $\mathbf{P}^\mu$.

Finally, it is straightforward to check that $\mathbf{S}^\mu$ and $\mathbf{A}^\mu$ are projections due to their idempotence, that is, $\mathbf{S}^\mu \mathbf{S}^\mu = \mathbf{S}^\mu$ and $\mathbf{A}^\mu \mathbf{A}^\mu = \mathbf{A}^\mu$. □

5.2 Proof of Theorem 4.6

Proof. We first consider the RKHS of the permutation invariant kernel $K^{PI}(v, v', \overline{v}, \overline{v}')$ given in Definition 3.2. According to the theorem concerning sums of reproducing kernels by Aronszajn [1950], the RKHS of the permutation invariant kernel K^{PI} can be written as the following space of functions:

$$
\begin{aligned}
\mathcal{H}\left(K^{PI}\right) &= \mathcal{H}\left(K + K^P\right) \\
&= \left\{ f_1 + f_2 : f_1 \in \mathcal{H}(K), f_2 \in \mathcal{H}\left(K^P\right) \right\} .
\end{aligned}
$$

This, together with the assumption $\mathcal{F} \subseteq \mathcal{H}(K)$, implies

$$
\mathcal{F} \subseteq \mathcal{H}\left(K^{PI}\right) . \tag{15}
$$

Let $\epsilon > 0$ and $t \in \mathcal{A}$ be an arbitrary function for which $t(v, v') = r(v, v') - r(v', v)$, where $r \in \mathcal{F}$. According to (15), we can select a set of pairs $\{(\overline{v}_i, \overline{v}'_i)\}_{i=1}^m$ and real numbers $\{\alpha_i\}_{i=1}^m$, such that the function

$$u(v, v') = \sum_{i=1}^{m} \alpha_i K^{PI}(v, v', \overline{v}_i, \overline{v}'_i)$$

belonging to the RKHS of the kernel K^{PI} fulfills

$$\max_{(v,v') \in \mathcal{P}^2} \{|r(v, v') - u(v, v')|\} \leq \frac{1}{2}\epsilon\,. \tag{16}$$

Let

$$h(v, v') = u(v, v') - u(v', v).$$

It follows from (16) that

$$\max_{(v,v') \in \mathcal{P}^2} \{|t(v, v') - h(v', v)|\} \leq \epsilon\,.$$

We observe that h can be written in terms of the kernel K^A as

$$\begin{aligned} h(v, v') &= \sum_{i=1}^{m} \alpha_i K^{PI}(v, v', \overline{v}_i, \overline{v}'_i) \\ &\quad - \sum_{i=1}^{m} \alpha_i K^{PI}(v', v, \overline{v}_i, \overline{v}'_i) \\ &= \sum_{i=1}^{m} \alpha_i K^{A}(v, v', \overline{v}_i, \overline{v}'_i) \end{aligned}$$

which proves the claim for the anti-symmetric kernels. The proof for the symmetric ones is analogous. □

5.3 Proof of Lemma 4.8

Proof. Starting from the form given by Cucker and Smale [2002] and applying the Sherman-Morrison-Woodbury fomula for operators [Deng, 2011], we get

$$\begin{aligned} f^\lambda &= \mathbf{U}_K^* h^\lambda \\ &= \mathbf{U}_K^* (\mathbf{U}_K \mathbf{U}_K^* + \lambda \mathbf{I})^{-1} \mathbf{U}_K f \\ &= \mathbf{U}_K^* \mathbf{U}_K (\mathbf{U}_K^* \mathbf{U}_K + \lambda \mathbf{I})^{-1} f \\ &= \mathbf{T}_K (\mathbf{T}_K + \lambda \mathbf{I})^{-1} f \\ &= \mathbf{V} \mathbf{\Lambda} (\mathbf{\Lambda} + \lambda \mathbf{I})^{-1} \mathbf{V}^* f\,. \end{aligned}$$

The bias caused by regularization is the squared error between the regression function and $\mathbf{U}_K^* h^\lambda$

$$\begin{aligned}
\epsilon_{rg}(f, \mathbf{T}, \lambda) &= \int_{\mathcal{X}} \left(f(x) - \mathbf{U}_K^* h^\lambda(x)\right)^2 d\mu \\
&= \left\langle f - f^\lambda, f - f^\lambda \right\rangle \\
&= \left\langle f - \mathbf{T}(\mathbf{T} + \lambda \mathbf{I})^{-1} f, f - \mathbf{T}(\mathbf{T} + \lambda \mathbf{I})^{-1} f \right\rangle \\
&= \left\langle f, \mathbf{V}\left(\mathbf{I} - 2\mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I})^{-1} + \mathbf{\Lambda}^2(\mathbf{\Lambda} + \lambda \mathbf{I})^{-2}\right)\mathbf{V}^* f \right\rangle \\
&= \left\langle f, \mathbf{V}\left(\mathbf{I} - \mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I})^{-1}\right)^2 \mathbf{V}^* f \right\rangle \\
&= \lambda^2 \left\langle f, \mathbf{V}\left(\mathbf{\Lambda} + \lambda \mathbf{I}\right)^{-2} \mathbf{V}^* f \right\rangle \\
&= \lambda^2 \left\langle f, \left(\mathbf{T} + \lambda \mathbf{I}\right)^{-2} f \right\rangle \ ,
\end{aligned}$$

where the products are in $L^2(\mathcal{X}, \mu)$. □

5.4 Proof of Theorem 4.9

Proof. Let the regression function be symmetric, that is, it can be written as $f = \mathbf{S}f$. Then,

$$\begin{aligned}
f_{K^{PI}}^\lambda &= \mathbf{V}\mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I})^{-1}\mathbf{V}^* f \\
&= \mathbf{V}\mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I})^{-1}\mathbf{V}^* \mathbf{S} f \\
&= \mathbf{V}\mathbf{\Lambda}_{K^S}(\mathbf{\Lambda}_{K^S} + \lambda \mathbf{I})^{-1}\mathbf{V}^* f \\
&= f_{K^S}^\lambda \ ,
\end{aligned}$$

where the second last inequality is due to the and hence also the bias caused by regularization is the same for the kernels K^{PI} and K^S. The proof is analogous for the anti-symmetric case.

Let $\mathbf{M}$ be an operator for which $0 < \alpha \mathbf{I} \leq \mathbf{M} \leq \beta \mathbf{I}$, where α and β are, respectively, the smallest and largest eigenvalues of $\mathbf{T} + \lambda \mathbf{I}$. We first recollect some matrix inequalities we use in the proof.

Choi's inequality and Kadison's inequality (see e.g. Choi [1974]) indiate that if $\mathbf{M} > 0$ and Ψ is positive and unital linear map, then

$$\Psi(\mathbf{M}^{-1}) \geq \Psi(\mathbf{M})^{-1} \tag{17}$$

$$\Psi(\mathbf{M}^2) \geq \Psi(\mathbf{M})^2 \ . \tag{18}$$

Let $0 < \alpha \leq \mathbf{M} \leq \beta$ and Ψ be positive unital linear map, Marshall and Olkin [1990] proved the following operator Kantorovich type of inequality:

$$\Psi(\mathbf{M}^{-1}) \leq \frac{(\alpha + \beta)^2}{4\alpha\beta} \Psi(\mathbf{M})^{-1} \ . \tag{19}$$

According to the Löwner-Heinz Theorem (see e.g. Carlen [2010]), if $\mathbf{M}$ and $\mathbf{N}$ are operators and $\mathbf{M} \geq \mathbf{N} \geq 0$, then matrix inversion reverses the positive-definite order, that is,

$$\mathbf{M}^{-1} \leq \mathbf{N}^{-1} \tag{20}$$

Further, Fujii et al. [1997] proved the following Kantorovich type of inequality:

$$\frac{(\alpha+\beta)^2}{4\alpha\beta}\mathbf{M}^2 \geq \mathbf{N}^2 \tag{21}$$

Armed with the above matrix inequalities, we get the following combined results:

$$\begin{aligned}\Psi(\mathbf{M})^{-2} &\geq \Psi(\mathbf{M}^2)^{-1} \\ &\geq \frac{4\alpha^2\beta^2}{(\alpha^2+\beta^2)^2}\Psi(\mathbf{M}^{-2}) \ ,\end{aligned}$$

where the first inequality is due to combining (18) with (20), and the second inequality is due to (19).

$$\begin{aligned}\Psi(\mathbf{M})^{-2} &\leq \frac{(\alpha^{-1}+\beta^{-1})^2}{4\alpha^{-1}\beta^{-1}}\Psi(\mathbf{M}^{-1})^2 \\ &= \frac{(\alpha+\beta)^2}{4\alpha\beta}\Psi(\mathbf{M}^{-1})^2 \\ &\leq \frac{(\alpha+\beta)^2}{4\alpha\beta}\Psi(\mathbf{M}^{-2}) \ ,\end{aligned}$$

where the first inequality is due to combining the Choi's inequality (17) with the inequality (21), and the second inequality is due to the Kadison's inequality (18).

Let $\Psi(\mathbf{M}) = \mathbf{SMS} + \mathbf{AMA}$, which is a unital, positive and linear mapping on $\mathcal{B}(L^2(\mathcal{X},\mu))$. Then, we have $\mathbf{T}_{K^{PI}} + \lambda\mathbf{I} = \Psi(\mathbf{T}_K + \lambda\mathbf{I})$. Combining the above results, we get

$$\begin{aligned}\epsilon_{rg}(f, \mathbf{T}_{K^{PI}}, \lambda) &= \lambda^2 \left\langle f, (\mathbf{T}_{K^{PI}} + \lambda\mathbf{I})^{-2} f \right\rangle \\ &= \lambda^2 \left\langle f, \Psi(\mathbf{T}_K + \lambda\mathbf{I})^{-2} f \right\rangle \\ &\leq \lambda^2 \frac{(\alpha+\beta)^2}{4\alpha\beta} \left\langle f, \Psi\left((\mathbf{T}_K + \lambda\mathbf{I})^{-2}\right) f \right\rangle \\ &= \lambda^2 \frac{(\alpha+\beta)^2}{4\alpha\beta} \left\langle \mathbf{S}f, \Psi\left((\mathbf{T}_K + \lambda\mathbf{I})^{-2}\right) \mathbf{S}f \right\rangle \\ &= \lambda^2 \frac{(\alpha+\beta)^2}{4\alpha\beta} \left\langle f, (\mathbf{T}_K + \lambda\mathbf{I})^{-2} f \right\rangle \ ,\end{aligned}$$

where the second last equality is due to the assumption of the regression function being symmetric. The lower bound can be shown analogously.

The limit of the smallest eigenvalue of $\mathbf{T}$ is 0, and hence that of $\mathbf{T} + \lambda\mathbf{I}$ is λ. Moreover, due to $K \max_{(v,v')\in\mathcal{P}^2} K(v, v', v, v') = 1$, the largest eigenvalue of $\mathbf{T} + \lambda\mathbf{I}$ is at most $1 + \lambda$. The claimed relationship is obtained by substituting λ and $1 + \lambda$ to α and β. □

References

Aronszajn, N. (1948). Rayleigh-ritz and a. weinstein methods for approximation of eigenvalues: I. operations in a hilbert space. *Proceedings of the National Academy of Sciences*, 34(10):474–480.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68.

Ben-Hur, A. and Noble, W. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1:38–46.

Brunner, C., Fischer, A., Luig, K., and Thies, T. (2012). Pairwise support vector machines and their application to large scale problems. *Journal of Machine Learning Research*, 13(1):2279–2292.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Carlen, E. (2010). Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum. Arizona school of analysis with applications*, pages 73–140. American Mathematical Society (AMS), Providence, RI, USA.

Choi, M.-D. (1974). A schwarz inequality for positive linear maps on c^*-algebras. *Illinois Journal of Mathematics*, 18(4):565–574.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Deng, C. Y. (2011). A generalization of the sherman-morrison-woodbury formula. *Applied Mathematics and Computation*, 24(9):1561–1564.

Fujii, M., Izumino, S., Nakamoto, R., and Seo, Y. (1997). Operator inequalities related to cauchy-schwarz and hölder-mccarthy inequalities. *Nihonkai Mathematical Journal*, 8(2):117–122.

Herbrich, R., Graepel, T., and Obermayer, K. (2000). Large margin rank boundaries for ordinal regression. In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press.

Hsu, D., Kakade, S. M., and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600.

Kashima, H., Oyama, S., Yamanishi, Y., and Tsuda, K. (2009). On pairwise kernels: An efficient alternative and generalization analysis. In Theeramunkong, T., Kijsirikul, B., Cercone, N., and Ho, T.-B., editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 1030–1037. Springer Berlin Heidelberg.

Li, Y. and Busch, P. (2013). Von neumann entropy and majorization. *Journal of Mathematical Analysis and Applications*, 408(1):384–393.

Mari, A., Giovannetti, V., and Holevo, A. S. (2014). Quantum state majorization at the output of bosonic gaussian channels. *Nature Communications*, 5.

Marshall, A. W. and Olkin, I. (1990). Matrix versions of cauchy and kantorovich inequalities. *Aequationes mathematicae*, 40:89–93.

Mendelson, S. (2003). On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771.

Pahikkala, T., Waegeman, W., Tsivtsivadze, E., Salakoski, T., and Baets, B. D. (2010). Learning intransitive reciprocal relations with kernel methods. *European Journal of Operational Research*, 206(3):676–685.

Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93.

Vert, J.-P., Qiu, J., and Noble, W. (2007). A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(Suppl 10):S8.

Waegeman, W., Pahikkala, T., Airola, A., Salakoski, T., Stock, M., and De Baets, B. (2012). A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, 20(6):1090–1101.

Zhang, T. (2002). Effective dimension and generalization of kernel learning. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 454–461. MIT Press.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098.

Zimmer, R. J. (1990). *Essential Results of Functional Analysis*. Chicago Lectures in Mathematics. University of Chicago Press.